Relational Learning techniques for Document Image Understanding: Comparing Statistical and Logical approaches

Michelangelo Ceci Margherita Berardi Donato Malerba
Dipartimento di Informatica – Università degli Studi di Bari
via Orabona 4 - 70126 Bari
{ceci, berardi, malerba}@di.uniba.it

Abstract

In this paper, we evaluate and systematically compare two different (multi-)relational learning methods based on a statistical approach and a logical approach for the task of document image understanding. For a fair comparison, both methods are tested on the same real world dataset consisting of multipage articles published in an international journal. An analysis of pros and cons of both approaches is reported.

1. Introduction

The increasingly large amount of paper documents to be processed daily requires systems with abilities to catalog and organize these documents automatically on the basis of their contents. Functional capabilities like classifying, storing, retrieving, and reproducing documents, as well as extracting, browsing and retrieving information from a variety of documents are highly demanded. In this context, the use of document image understanding techniques to recognize semantically relevant layout components (e.g. *title*, *abstract* of a scientific paper or *leading article*, *picture* of a newspaper) in the layout extracted from a document image plays a key role.

This recognition process is based on some visual models, whose manual specification can be a highly demanding task. In order to automatically acquire these models, machine learning methods characterized by a high degree of adaptivity can be used. In the literature, several machine learning techniques have been applied for the document image understanding task. Aiello et al. [1] applied the classical decision tree learning system C4.5 [13] to learn classification rules for recognizing textual layout components. Palmero et al. [10] developed a neuro-fuzzy learning algorithm that ranks, for each new (unseen) block, candidate labels and selects the best. Le Bourgeois et al. [7] proposed to use the probabilistic relaxation [14] and Bayesian Networks [11] for recognizing logical components. Walischewski [17] proposed to represent each document layout by a complete attributed directed graph (one vertex for each layout object) that represents frequency counts for different spatial relations. Incremental learning of the attributed directed graph is proposed. Akilende el al. [2] proposed to infer a tree-based representation of the layout structure by means of a tree-grammar inference method from training documents.

Although these methods often present interesting results, they are often based on learning algorithms that suffer from severe limitations due to the restrictive representation formalism known as single-table assumption [6]. More specifically, it is assumed that training data are represented in a single table of a relational database, such that each row (or tuple) represents an independent example and columns correspond to properties. This requires that non-spatial properties of neighboring objects be represented in aggregated form causing a consequent loss of information. On the contrary, the application of (multi-)relational learning techniques [6] allows spatial relations between layout components to be effectively and naturally represented, while, for example, decision trees and neural networks models are unsuitable to represent a variable number of spatial neighbours of a layout component together with their attributes.

Scope of this paper is to evaluate and systematically compare two different (multi-)relational learning approaches for document image understanding, based on statistical approaches and logical approaches, respectively. In particular, we consider the statistical learner Mr-SBC [4], and the logical learner ATRE [8].

In order to test Mr-SBC and ATRE for the document image understanding task, both have been integrated in the Document Image Analysis system WISDOM++ (www.di.uniba.it/~malerba/wisdom++/) [3] whose applicability has been investigated in the context of the IST-EU founded project COLLATE (www.collate.de/). WISDOM++ permits the transformation of document images into XML format by means of several complex steps: preprocessing of the raster image of a scanned paper document, segmentation of the preprocessed raster image into basic layout components, classification of basic layout components according to the type of content (e.g., text, graphics, etc.), the identification of a more abstract representation of the document layout (layout analysis), the classification of the document on the ground of its layout and content, the identification of semantically relevant layout components (document image understanding), the application of OCR to textual components of interest, the storage in XML format.

The paper is organized as follows. In section 2 and 3, the application of the learning systems Mr-SBC and ATRE in the context of document image understanding is described. In section 4 experimental results on the same dataset consisting of real world documents are shown and conclusions are drawn.

2. Application of Mr-SBC

Mr-SBC (Multi-Relational Structural Bayesian Classifier) [4] is a (multi-)relational classifier that combines the induction of first order logic classification rules and the classical naive bayesian classifier [5]. In particular, it can be considered an extension of the naive Bayesian classifier in the case of the multi relational setting.

Mr-SBC is particularly suited in the task in hand since it is tightly-coupled with a Relational DBMS and can directly interface, by means of SQL views, the database that WISDOM++ uses for storing intermediate data. Mr-SBC takes advantage of the database schema that provides useful knowledge of data model that can help to guide the learning process. This is an alternative to asking the users to specify background knowledge.

The problem solved by Mr-SBC can be formalized as follows: *Given*:

- a training set represented by means of h relational tables $S=\{T_0,T_1,\ldots,T_{h-1}\}$ of a relational database D.
- a set of primary key constraints on tables in S.
- a set of foreign key constraints on tables in S.
- a target relation $T\in S$
- a target discrete attribute y in T.

Find a naive Bayesian classifier which predicts the value of y for some individual represented as a tuple in T (with possibly UNKNOWN value for y) and related tuples in S according to foreign key constraints.

According to the Bayesian setting, given a new instance to be classified, the classifier estimates the probability that an instance belongs to a determinate class and returns the most probable class:

$$f(I)= arg\ max_i\ P(C_i|R) = arg\ max_i\ \frac{P(C_i)P(R|C_i)}{P(R)}$$

where $f(\cdot)$ is the classification function, I is the individual to be classified, C_i is the i-th possible class and R is the description of I in terms of first order classification rules. In our domain, categories are logical labels that can be associated to layout components (individuals).

Although Mr-SBC can be used for Document Image Understanding tasks, some modifications are necessary. In particular, it is necessary to modify the search strategy in order to allow cyclic paths. As observed by Taskar et al. [16], the acyclicity constraint hinders representation of many important relational dependencies. This is particularly true in the task in hand, where a relation between two logical components is modelled by means of a relational table that expresses the existence of the topological relation. For example, suppose that we need to model the relation *on_top* between two layout components, from a database point of view, this is realized by means of the table "block" and a table "on_top" that contains two foreign keys to the table block. The referenced blocks are considered one on top the other. In the original formulation of the problem solved by Mr-SBC, first order classification rules do not consider the same table twice [4], therefore it is not possible to take into account the topological relation. To avoid this problem, we modified Mr-SBC, allowing cyclic paths.

The second problem concerns with the classification of layout components. In document image understanding, it is possible that the same layout component is associated with two different logical labels. For example, suppose that the layout analysis is not able to separate the page number and the running head of a scientific paper. In this case we have a single layout component that contains two logical components: the page number and the running head. The classifier should associate that component with two labels. For this reason, it is necessary to resort to a multiple classification problem. In particular, we learn a binary classifier for each class. Each classifier is able to identify examples belonging to that class and examples that do not belong to it. This solution is usually adopted in Text Categorization when the problem is to establish if a document belongs to a particular class or not [15].

However, the use of multiple classification leads to the problem of "unbalanced datasets". In fact, data can be characterized by a predominant number of negative examples with respect to the number of positive examples (e.g. in the examples reported in section 4, the percentage of layout components classified as "*table*" is 1.4% of all components). Several approaches that face the problem of the unbalanced datasets have been proposed in the literature. Some of them are based on a sampling of examples in order to have a balanced dataset [9]. A different approach is: given the class, a ranking of all the examples in the test set from the most probable member to the least probable member is computed and then, a correctly calibrated estimate of the true probability that each test example is a member of the class of interest is computed [18]. In other words, a probability threshold that delimitates the membership and the non-membership of a given test example to the class is computed. In our approach, we exploit the consideration that the naive

Bayesian classifier for two-class problems tends to rank examples well (even if the classifier does not return a correct probability estimation)[18]. In our solution, the threshold is determined by maximizing the AUC (Area Under the ROC Curve) [12] according to a cost function:

$$cost = \mathrm{P}(C_i)\cdot(1\text{-}TP)\cdot \mathrm{c}(\neg C_i;\, C_i) + \mathrm{P}(\neg C_i)\cdot FP\cdot \mathrm{c}(C_i;\, \neg\, C_i)$$

where $\mathrm{P}(C_i)$ is the a-priori probability that an example belongs to the class C_i, $\mathrm{P}(\neg C_i)$ is the a-priori probability that an example does not belong to the class C_i, $\mathrm{c}(\neg C_i;\, C_i)$ is the cost of classifying a positive example as negative (for the class C_i) and $\mathrm{c}(C_i;\, \neg C_i)$ is the cost of classifying a negative example as positive. *TP* is the true positive rate and *FP* is the false positive rate. We denote as *CostRatio* the value: $CostRatio = \mathrm{c}(C_i;\, \neg C_i)/\mathrm{c}(\neg C_i;\, C_i)$.

The Mr-SBC database input schema (see figure 1) represents the logical structure of a document image. In particular, we represent locational features (i.e. *x_pos_center* and *y_pos_center*, that is, position of the component with respect to a coordinate system), geometrical features (i.e. *width* and *height*), topological relations (i.e. *on_top*, *to_right*, *only_right_col*, *only_middle_row*, *only_lower_row*, *only_middle_col*, *only_left_col* and *only_upper_row*) and aspatial features (i.e. *type_of* that specifies the content type of a logical component, that is, image, text, horizontal line).

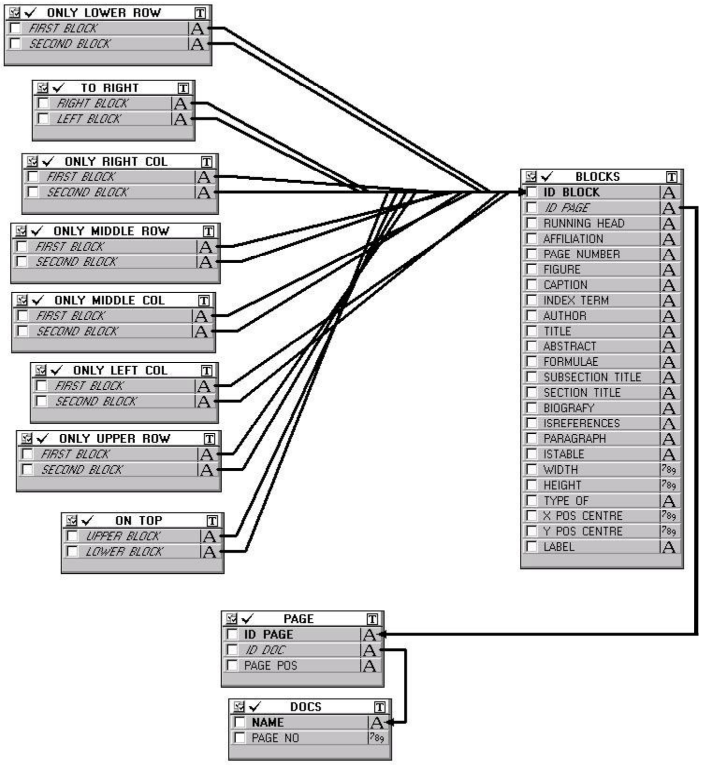

Fig. 1. Mr-SBC Database input schema

3 Application of ATRE

ATRE [8] is a (multi-) relational ILP (Inductive Logic Programming) system that can learn logic theories from examples and which is able to handle symbolic as well as numerical descriptors. In this framework, ATRE learns first order rules that can be subsequently used in the classification step. Formally, ATRE solves the following learning problem: *Given:*

- a set of concepts $C_1, C_2, \ldots, C_r$ to be learned,
- a set of observations O described in a language L_O,
- a background knowledge expressed in a language L_{BK},
- a language of hypotheses L_H,
- a generalization model Γ over the space of hypotheses,
- a user's preference criterion *PC*,

Find a (possibly recursive) logical theory T for the concepts $C_1, C_2, \ldots, C_r$, such that T is complete and consistent with respect to O and satisfies *PC*.

The completeness property holds when the theory T explains all observations in O of the r concepts C_i, while the *consistency* property holds when the theory T explains no counter-example in O of any concept C_i. The satisfaction of these properties guarantees the correctness of the induced theory with respect to O.

In ATRE, observations are represented by means of ground multiple-head clauses, called *objects*. In this application, each object corresponds to a document page. All literals in the head of the clause are called *examples* of the concepts $C_1, C_2, \ldots, C_r$. They can be considered either positive or negative according to the learning goal. In this application domain, the concepts to be learned are logical labels (e.g. *title(X)=true*, *page_number(X)=true*, etc.), since we are interested in finding rules which predict the logical label of a layout component. No rule is generated for the case *title(X)=false*.

The generalization model provides the basis for organizing the search space, since it establishes when a hypothesis explains a positive/negative example and when a hypothesis is more general/specific than another. The generalization model adopted by ATRE, called *generalized implication*, is explained in [8].

The preference criterion *PC* is a set of conditions used to discard/favour some solutions. In this work, short rules, which explain a high number of positive examples and a low number of negative examples, are preferred.

In ATRE, the application of a first-order logic language permits to represent both unary (*attributes*) and binary function symbols (*relations*). Attributes are used to describe properties of a single layout component (e.g., *height*), while relations are used to express spatial relationships among layout components (e.g., *part-of* or *on_top*). Similarly to the case of Mr-SBC, for ATRE the following descriptors have been used: *width(block)*, *height(block)*, *x_pos_centre(block)*, *y_pos_centre(bockl)*, *type_of(block)*, *part_of(page,block)*, *on_top(boclk1, blk2)*, *to_right(block1,block2)*, *alignment(block1, block2)*.
An example of an object representation follows:

```
class(1)=.tpami,
running_head(2)=.true, ..., table(2)=.false,
title(3)=.true,..., table(3)=.false,...
←page(1)=.first, part_of(1,2)=.true,...,
   width(2)=.390,...,on_top(2,3)=.true,...,
   alignment(7,8)=.only_middle_row.
```

where the constant 1 denotes the whole page, while the constants 2, 3, …,15 denote the layout components.

4 Experiments

For a fair comparison of the two learning methods, both Mr-SBC and ATRE are trained on the same dataset consisting of twenty-one articles published as either regular or short in the IEEE TPAMI. Each paper is a multi-page document; therefore, we processed 197 document images in all and the user manually labeled 2436 layout components, that is, in average, 116 components per document, 12.37 per page. About 74% of the layout components have been labeled, the remaining components are "irrelevant" for the task in hand or are associated to "noise" blocks: they are automatically considered *undefined*. A description of the dataset is reported in table 1.

The performance of the learning tasks is evaluated by means of a 5-fold cross-validation, that is, the set of twenty-one documents is first divided into five folds, and then, for every fold, Mr-SBC and ATRE are trained on the remaining folds and tested on the hold-out fold. In the case of Mr-SBC, we set *CostRatio*=10 since, for that value, the system showed the best performances among 11 trials (*CostRatio*={1,2,4,...,20}). Due to space limitations we report results only for *CostRatio*=10.

For each learning problem, the number of omission/commission errors is recorded. Omission errors occur when logical labelling of layout components are missed, while commission errors occur when wrong logical labelling are "recommended" by classifiers. In our study we do not consider the standard classification accuracy, because for each learning task, the number of positive and negative examples is strongly unbalanced and, in most cases, the trivial classifier that returns always "undefined" would be the classifier with the best accuracy. On the contrary, we are generally interested in reducing omission errs rather than maximizing accuracy.

In table 2 results are reported and permit to compare the two systems both in terms of efficiency and effectiveness of the learning task. We note that the statistical classifier is, in general, more efficient than the logical approach. In terms of omission errors, the two systems do not show great difference. However, looking at results on commission errors, we can conclude that ATRE outperforms Mr-SBC in terms of classification effectiveness. In a deeper analysis, we note that Mr-SBC outperforms ATRE, in terms of omission errors, when the size of the layout component does not show great variability (e.g. this is the case of *section title*, *subsection title*, *title* and is not the case of *table* and *figure*). This can be explained by considering that the discretization algorithm implemented in Mr-SBC does not take into account combination of features (e.g. the size of a layout component is computed independently from the page order)[4]. This aspect negatively affects the learned classification model. Concerning ATRE, we note that the results are characterized by a high percentage of omission

Table. 1. Dataset description: Distribution of pages and examples per document grouped by 5 folds.

Fold No	Document name	No. of pages	No of labeled components	Tot No. of components
1	TPAMI_1	13	476	597
	TPAMI_13	3		
	TPAMI_14	10		
	TPAMI_16	14		
2	TPAMI_8	5	519	684
	TPAMI_15	15		
	TPAMI_18	10		
	TPAMI_24	6		
3	TPAMI_3	15	481	697
	TPAMI_7	6		
	TPAMI_12	6		
	TPAMI_20	14		
4	TPAMI_9	5	541	774
	TPAMI_11	6		
	TPAMI_19	20		
	TPAMI_21	11		
5	TPAMI_4	14	419	549
	TPAMI_6	1		
	TPAMI_10	3		
	TPAMI_17	13		
	TPAMI_23	7		
Tot.	21 docs	197	2436	3301

Table. 2. Mr-SBC vs. ATRE: Average number of omission errors over positive examples, commission errors over negative examples and learning times (in secs).

	Omiss/Pos		Comm/Neg		Learning Times (s)	
	ATRE	Mr-SBC	ATRE	Mr-SBC	ATRE	Mr-SBC
Abstract	**0.55**	0.81	0.00	0.21	660	492
Affiliation	**0.50**	0.77	0.00	0.25	756	564
Author	0.46	**0.40**	0.00	0.26	732	504
Biography	0.63	**0.57**	0.00	0.26	636	444
Caption	**0.68**	0.74	0.03	0.23	12240	552
Figure	**0.13**	0.62	0.02	0.23	4440	960
Formulae	**0.45**	0.57	0.06	0.07	21120	624
Index Term	0.53	**0.27**	0.00	0.22	169.6	564
Reference	0.95	**0.60**	0.01	0.21	1884	480
Table	**0.69**	0.83	0.01	0.06	1668	528
Page No	**0.04**	0.26	0.00	0.01	490	660
Paragraph	----	**0.89**	----	0.03	----	1572
RunningHea	**0.09**	0.55	0.00	0.01	485.4	504
Section Title	0.80	**0.48**	0.01	0.27	2052	516
Subsect.Titl.	1.00	**0.72**	0.00	0.27	1068	468
Title	0.60	**0.39**	0.00	0.25	648	636

errors and a low percentage of commission errors. This is due to a lower percentage of positive examples that generally leads to a specificity of learned rules with a low percentage of coverage of training examples. Specificity of learned rules is due to the fact that ATRE is asked to generate a complete theory, that is a set of rules that explain all positive examples. Moreover, ATRE was not able to learn the concept "*paragraph*" since the high number of positive examples significantly increases the complexity of the task.

For a complete analysis, we have to consider that the statistical classifier is able to rank the layout components giving a "confidence" of the classification. Such information can help the user to manually correct and interpret classification results. On the other hand, ATRE returns a set of first order rules that can be easily interpreted by the user, thus allowing to understand the decisions taken. An example of rule learned by ATRE is:

```
abstract(X1)= true←
alignment(X1,X2)=  only_right_col,  height(X2)in
[384...422], y_pos_centre(X1)in[169...197].
```

This rule states that a layout component with the baricentre at a point between 169 and 197 on the y-axis and vertically aligned with a layout component with height between 384 and 422 (e.g. the title) is an abstract.

5 Conclusions

This work presents an application of (multi-)relational learning techniques to the problem of document image understanding. In particular, two learning methods, namely Mr-SBC and ATRE, based on a statistical approach and a logical approach, respectively, are compared. For the evaluation, both methods have been embedded in the DIA system WISDOM++.

While Mr-SBC directly interfaces the internal WISDOM++ database schema, ATRE needs some preprocessing in order to transform the internal representation of the layout structure in first order logic representation. Results show that Mr-SBC is more efficient than ATRE in terms of running times. Concerning classification effectiveness, while ATRE outperforms Mr-SBC in terms of omission errors, in terms of commission errors the systems do not show great differences. Weaknesses for Mr-SBC and ATRE are respectively due to the discretization algorithm and to the strong assumptions of completeness and consistency of the learned theories. In terms of understandability of the learned model, although Mr-SBC provides a confidence of the classification result, ATRE provides a set of rules that are easily comprehensible to humans.

For future work, we intend to improve the Mr-SBC algorithm allowing contextual discretization and to explore the opportunity of weakening the conditions of applicability of rules in ATRE in order to significantly recover omission errors.

Acknowledgments

The work presented in this paper is partial fulfillment of the research objective set by the ATENEO-2004 project on "Metodi di Data Mining Multi-relazionale per la scoperta di conoscenza in basi di dati".

References

1. Aiello M., Monz C., Todoran L., Worring M.: Document Understanding for a Broad Class of Documents. *IJDAR* 5(1), 1-16, 2002.
2. Akindele O.T., Belaïd A.: Construction of generic models of document structures using inference of tree grammars, *Proc. of the 3rd ICDAR*, 206-209, 1995
3. Altamura O., Esposito F., & Malerba D.: Transforming paper documents into XML format with WISDOM++, *IJDAR*, 4(1), 2-17, 2001.
4. Ceci M., Appice A., Malerba D.: Mr-SBC: a Multi-Relational Naive Bayes Classifier. *In 7th European Conference PKDD '03* LNAI 2838: 95–106, 2003
5. Domingos P. and Pazzani M.. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.
6. Dzeroski S., Lavrac N.: Relational Data Mining. Springer-Verlag, Berlin Germany, 2001.
7. Le Bourgeois F., Souafi-Bensafi S., Duong J., Parizeau M., Coté M., Emptoz H. Using statistical models in document images understanding. *In proc.Workshop DLIA '01*, 2001.
8. Malerba D.: Learning recursive theories in the normal ilp setting. Fundamenta Informaticae 57(1):39–77. 2003
9. Mladenic D. and Grobelnik M.: Feature selection for unbalanced class distribution and naive bayes. *Int. Conf. on Machine Learning ICML'99*, 258–267. 1999.
10. Palmero G.I.S., Dimitriadis Y.A.: Structured Document Labeling and Rule Extraction using a New Recurrent Fuzzy-neural System. *ICDAR*. 181-184, 1999.
11. Pearl J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, 1988.
12. Provost F., Fawcett T.: Robust classification for imprecise environments. *Machine Learning*. 42(3):203–231, 2001.
13. Quinlan J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., 1993.
14. Rosenfeld A., Hummel R.A., Zucker S.W.: Scene labeling by relaxation operations, *IEEE trans SMC* 6(6), 1976.
15. Sebastiani F.: Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1-47, 2002.
16. Taskar B., Abbeel P., and Koller D.: Discriminative probabilistic models for relational data. *Proc. of Int. Conf. on Uncertainty in Artificial Intelligence*, 485-492, 2002.
17. Walischewski H.: Automatic knowledge acquisition for spatial document interpretation. *ICDAR*, 243-247, 1997.
18. Zadrozny B. and Elkan C.: Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *Int. Conf. on Machine Learning* 609–616, 2001.